CS 696 Applied Large Language Models
Spring Semester, 2025
Doc 8 Reducing Heads
Feb 6, 2025

Are Sixteen Heads Really Better than One?
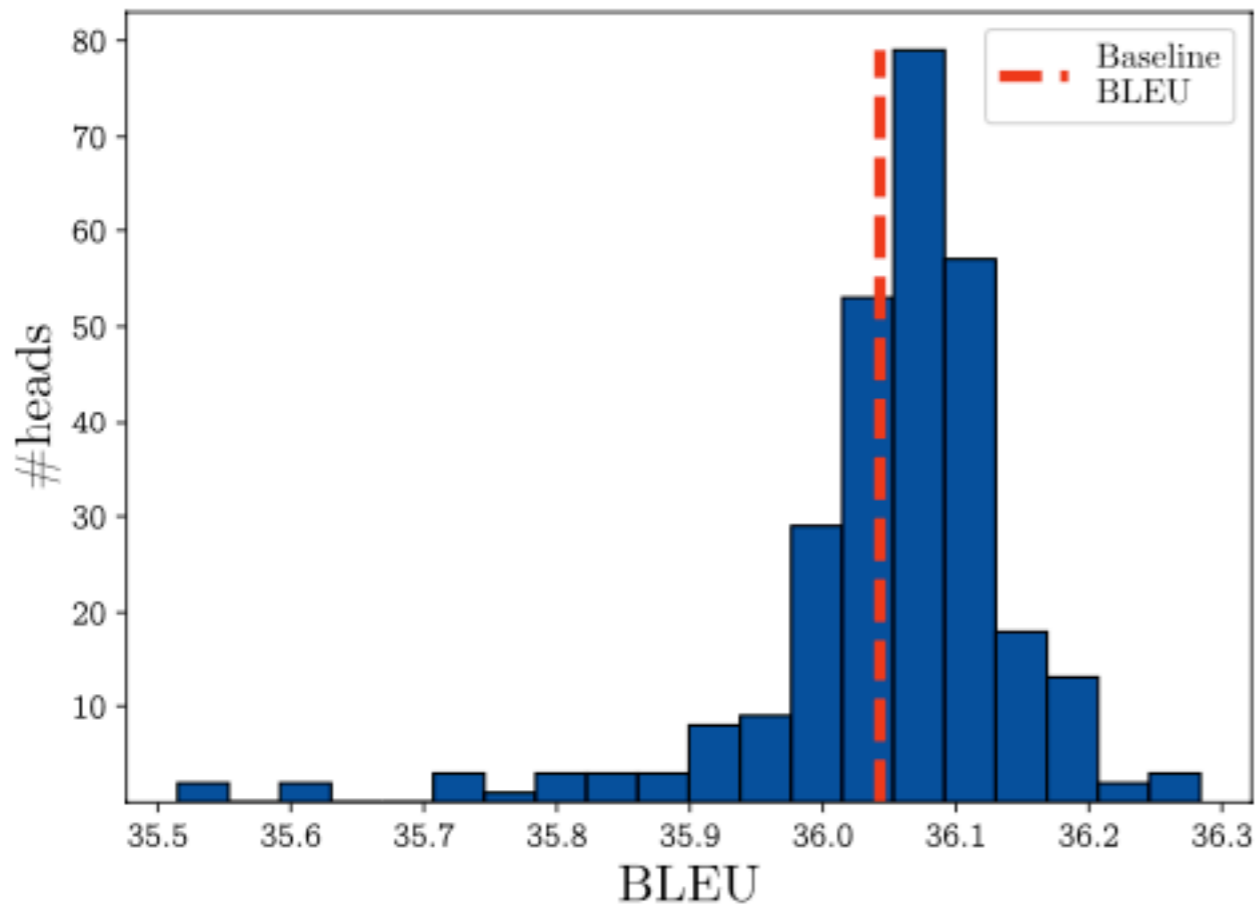
(a) WMT
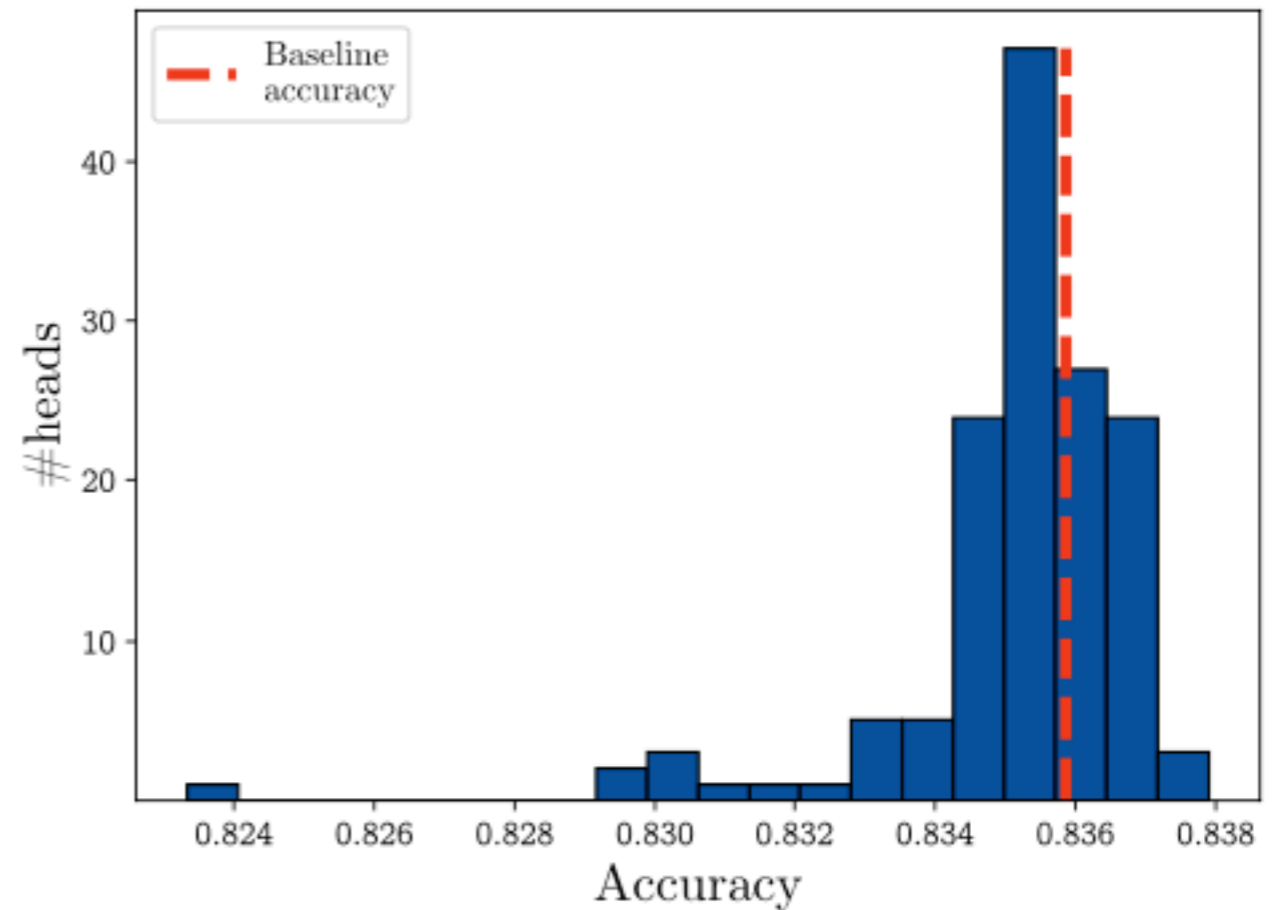
(b) BERT

Figure 1: Distribution of heads by model score after masking.

Ablating One Head

Head Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.03	0.07	0.05	-0.06	0.03	**-0.53**	0.09	**-0.33**	0.06	0.03	0.11	0.04	0.01	-0.04	0.04	0.00
2	0.01	0.04	0.10	**0.20**	0.06	0.03	0.00	0.09	0.10	0.04	**0.15**	0.03	0.05	0.04	0.14	0.04
3	0.05	-0.01	0.08	0.09	0.11	0.02	0.03	0.03	-0.00	0.13	0.09	0.09	-0.11	**0.24**	0.07	-0.04
4	-0.02	0.03	0.13	0.06	-0.05	0.13	0.14	0.05	0.02	0.14	0.05	0.06	0.03	-0.06	-0.10	-0.06
5	**-0.31**	-0.11	-0.04	0.12	0.10	0.02	0.09	0.08	0.04	**0.21**	-0.02	0.02	-0.03	-0.04	0.07	-0.02
6	0.06	0.07	**-0.31**	0.15	-0.19	0.15	0.11	0.05	0.01	-0.08	0.06	0.01	0.01	0.02	0.07	0.05

Iterative Pruning of Attention Heads
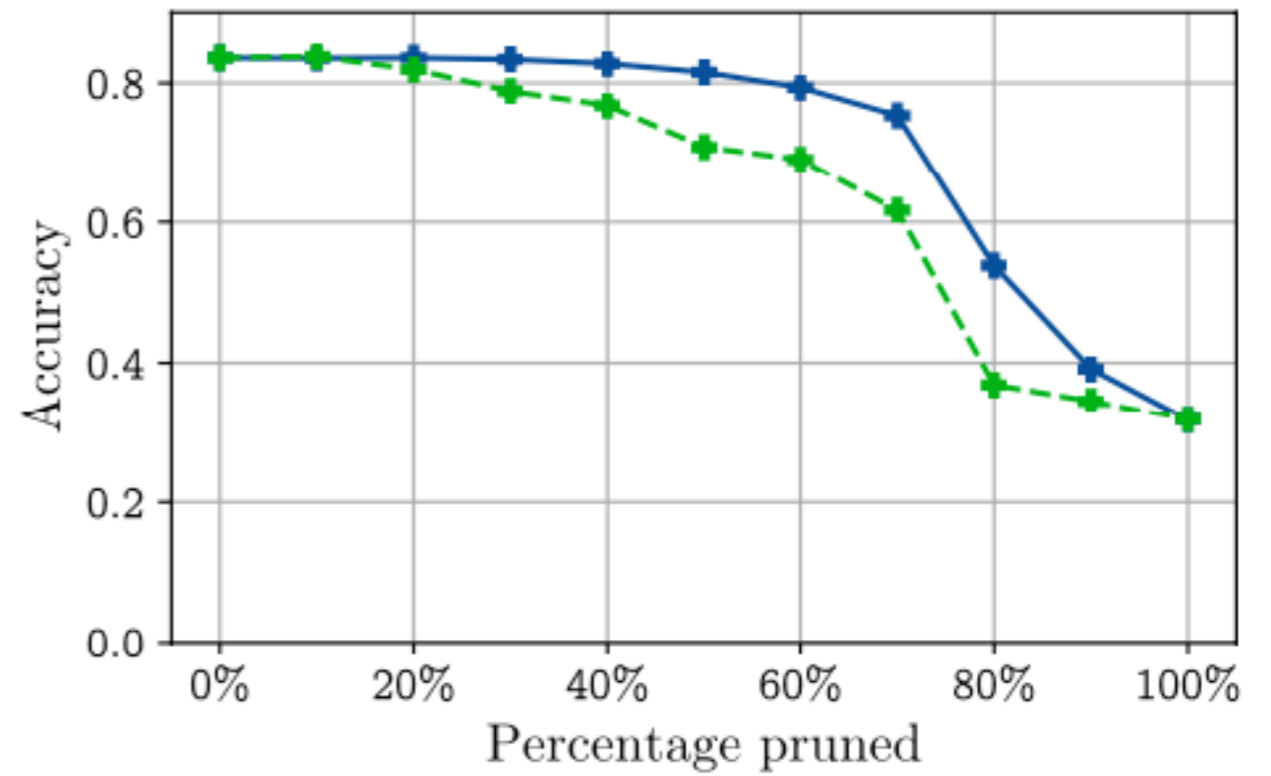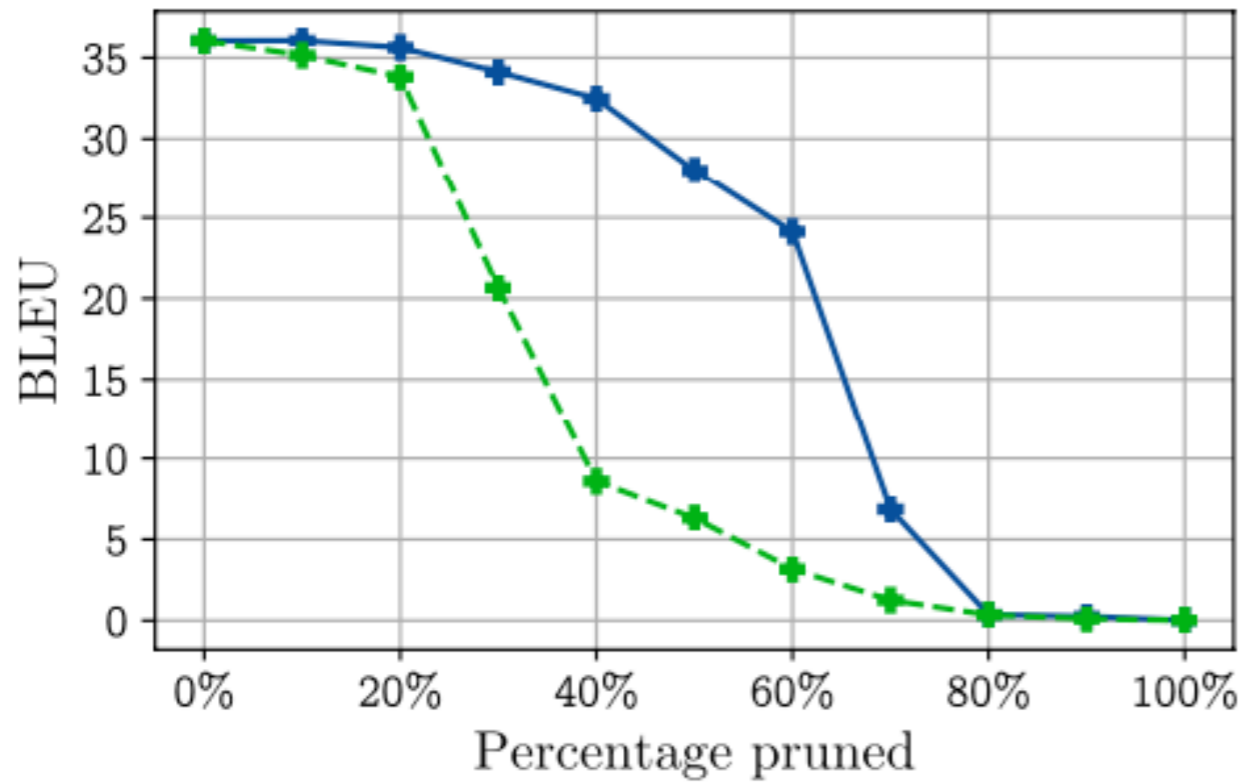
When Are More Heads Important
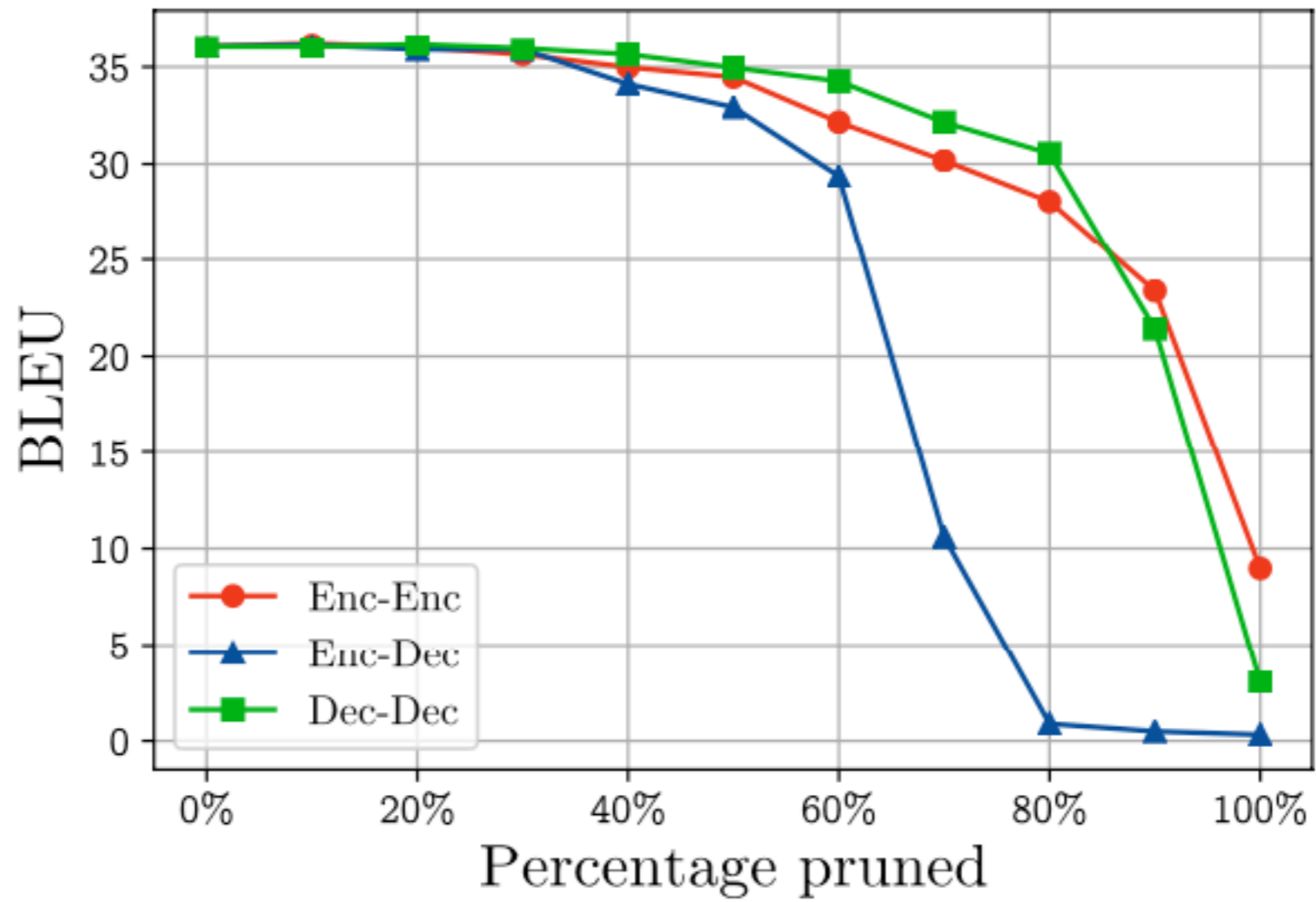
BERTology

See BERTology on Huggingface for how to access details of BERT and measure
Some performance, but will not run on GPU Cluster